J Mol Model (2016) 22: 49
DOI 10.1007/s00894-016-2915-2

ORIGINAL PAPER

A fast loop-closure algorithm to accelerate residue matching in computational enzyme design

Jing Xue[1] · Xiaoqiang Huang[1] · Min Lin[1] · Yushan Zhu[1]

Received: 2 November 2015 / Accepted: 11 January 2016 / Published online: 29 January 2016

Abstract Constructing an active site on an inert scaffold is still a challenge in chemical biology. Herein, we describe the incorporation of a Newton-direction-based fast loop-closure algorithm for catalytic residue matching into our enzyme design program ProdaMatch. This was developed to determine the sites and geometries of the catalytic residues as well as the position of the transition state with high accuracy in order to satisfy the geometric constraints on the interactions between catalytic residues and the transition state. Loop-closure results for 64,827 initial loops derived from 21 loops in the test set showed that 99.51 % of the initial loops closed to within 0.05 Å in fewer than 400 iteration steps, while the large majority of the initial loops closed within 100 iteration steps. The revised version of ProdaMatch containing the novel loop-closure algorithm identified all native matches for ten scaffolds in the native active-site recapitulation test. Its high speed and accuracy when matching catalytic residues with a scaffold make this version of ProdaMatch potentially useful for scaffold selection through the incorporation of more complex theoretical enzyme models which may yield higher initial activities in de novo enzyme design.

Keywords Computational enzyme design · Protein design · Loop closure algorithm · Numerical optimization

Electronic supplementary material The online version of this article (doi:10.1007/s00894-016-2915-2) contains supplementary material, which is available to authorized users.

✉ Yushan Zhu
yszhu@tsinghua.edu.cn

[1] Department of Chemical Engineering, Tsinghua University, Beijing 100084, People's Republic of China

Introduction

Over the course of evolution, living systems have developed native enzymes as a means to catalyze the chemical reactions that are required for their survival and reproduction [1]. Some of the most striking features of enzymes as biocatalysts are that they show high catalytic efficiency often at or near the diffusion limit, and they catalyze reactions in a manner that leads to enormous rate enhancements in water at neutral pH values and mild temperatures [2]. These properties have led to the use of enzymes by the pharmaceutical and chemical industries to catalyze a great variety of chemical transformations in environmentally benign synthetic processes [3]. However, the high catalytic efficiencies of enzymes are always coupled to their exceptional chemo-, regio-, and stereoselectivities, limiting their usefulness for catalyzing reactions with non-native substrates. To overcome this obstacle, protein engineering approaches such as directed evolution and rational design have been extensively used to broaden the scope of enzyme-catalyzed chemical transformations [4–6]. Random mutagenesis-based directed evolution technology has been successfully employed to develop efficient enzymes for alternative substrates starting from those with very weak but nonselective activities. However, it is more difficult to develop enzymes for nonbiological reactions that do not have a natural counterpart. Therefore, experimental screening approaches for evolving enzymes that are capable of catalyzing synthetic substrates need to start with substrates that possess activities above background levels. Structure-based computational enzyme design technology is currently being rapidly developed to meet this challenge [7, 8]. In contrast with catalytic antibody technology, de novo enzyme design is not limited to the immunoglobulin scaffold; it can utilize the set of all known protein scaffolds, many of which have proven tractable for further optimization of enzyme activity using either directed evolution or computational redesign [9].

The computational enzyme design process can be roughly divided into four steps [10]: (i) construction of a theozyme model for the target reaction [11]; (ii) matching catalytic residues with potential scaffolds from a scaffold library; (iii) selection of suitable amino acids at the binding sites; and (iv) experimental validation. The second step is critical to the successful design of an active enzyme for the target reaction, and several programs have been developed to fulfill this task, including Dezymer [12], Orbit [13], RosettaMatch [14], OptGraft [15], ScaffoldSelection [16], ProdaMatch [17], AutoMatch [18], and Saber [19]. RosettaMatch was experimentally shown to create de novo enzyme catalysts for reactions including the Kemp elimination [20], the retro-aldol reaction [21], the Diels–Alder reaction [22], hydrolytic reactions of esters [23], and the Morita–Baylis–Hillman reaction [24]. Orbit was used to identify enzyme-like proteins for hydrolytic reactions of esters [25] as well as de novo enzymes for the Kemp elimination reaction [26].

ProdaMatch [17] was recently developed to match multiple residues with a scaffold in the expectation that the complex theozyme geometries of these residues might lead to higher starting activities in the more demanding reactions. ProdaMatch is a rotamer-library-free approach where the catalytic residues can adopt unusual high-energy-motif geometries which may be required for enzyme catalysis but are not included in rotamer libraries because they always comprise low-energy conformers of naturally occurring amino acids. The geometries of the catalytic residues and the catalytic relative geometries of the catalytic residues and the transition state (TS) are determined by the loop-closure algorithm. The use of a continuous optimization-based algorithm can effectively curb the combination explosion caused by the discrete sampling of these geometries. Also, the loop-closure problem in our former implementation of ProdaMatch can be solved via an extended cyclic coordinate descent (CCD) algorithm. CCD is an iterative approach that is based on univariate optimization and is inefficient, as just one direction is adjusted in each step, though it is also simple and easy to implement. In the matching process, thousands of loop-closure problems need to be solved quickly, which means that this matching process must be as efficient as possible. Given the inefficiency of the CCD algorithm, part of the tradeoff for achieving optimal matching efficiency is that inaccurate loop closure will sometimes occur. Additionally, inaccurate loop closure may destroy the catalytic relative geometries of the catalytic residues and the TS. While this problem is not appreciable for benchmark tests on native scaffolds, it becomes more pronounced for de novo design on inert scaffolds with complex theozyme models. Moreover, similar to loop closure for protein structure prediction [27], the CCD algorithm for enzyme design favors large changes in the variables, and the optimized bond lengths and angles are always at the boundaries of their allowed ranges, leading to considerable deviations from the perfect geometries. Another popular loop closure method, CSJD [28, 29], which is based upon an analytical root-finding algorithm and outperforms CCD in terms of accuracy and speed, may not readily converge, as many bounded bond-length and bond-angle variables must be optimized besides the unbounded torsion angles in loop-closure problems associated with enzyme design. To tackle these problems, a quasi-Newton-direction-based fast loop-closure algorithm was developed to shorten the time required for highly accurate residue matching in ProdaMatch during computational enzyme design.

Materials and methods

Given the constraints on the catalytic relative geometriies of the catalytic residues and the transition state in the target reaction, the catalytic residue site selection problem in enzyme design can be defined as a matching process that anchors the catalytic residues onto the potential sites in the active pocket of a protein scaffold. Assuming that the number of catalytic residues is m, and that there are n sites in the active pocket to which those residues can attach, the total number of combinations for this matching problem is $n \times (n-1) \times \ldots \times (n-m+1)$. When the number of catalytic residues is small, this problem is tractable, even when performing an exhaustive search. The matching algorithm, ProdaMatch, developed in our earlier work [17], enumerates all of the combinations sequentially. After the sites have been putatively specified for all catalytic residues, the conformations of the catalytic residues as well as the catalytic relative geometries of the catalytic residues and the transition state are determined by the continuous optimization approach. Each catalytic geometrical relationship between species can be described by one bond length, two angles, and three torsion angles (see Fig. 1 in [17]). If the continuous optimization approach can identify allowable conformations for all catalytic residues and suitable catalytic geometrical relationships between the transition state and catalytic residues, this particular combination will represent a feasible match for the target reaction. Since discrete sampling of conformations of catalytic residues and catalytic geometrical relationships is avoided through the application of the continuous optimization approach, the total number of combinations tested in ProdaMatch is effectively reduced. After all of the catalytic residues have been assigned to their corresponding sites during the enumeration process, the continuous optimization process is initiated by anchoring atoms N, CA, and CB of the first catalytic residue onto the corresponding positions of atoms N, CA, and CB at its candidate site. A loop is formed from the atom CA of the first catalytic residue to the atom CA of the second catalytic residue via the transition state, and atoms CB and CA of the second catalytic residue are allowed to move. Therefore, the goal of the continuous optimization approach is to adjust the side-chain torsion angles of the two catalytic residues and the parameters associated with the catalytic geometrical relationships, such as bond lengths, bond

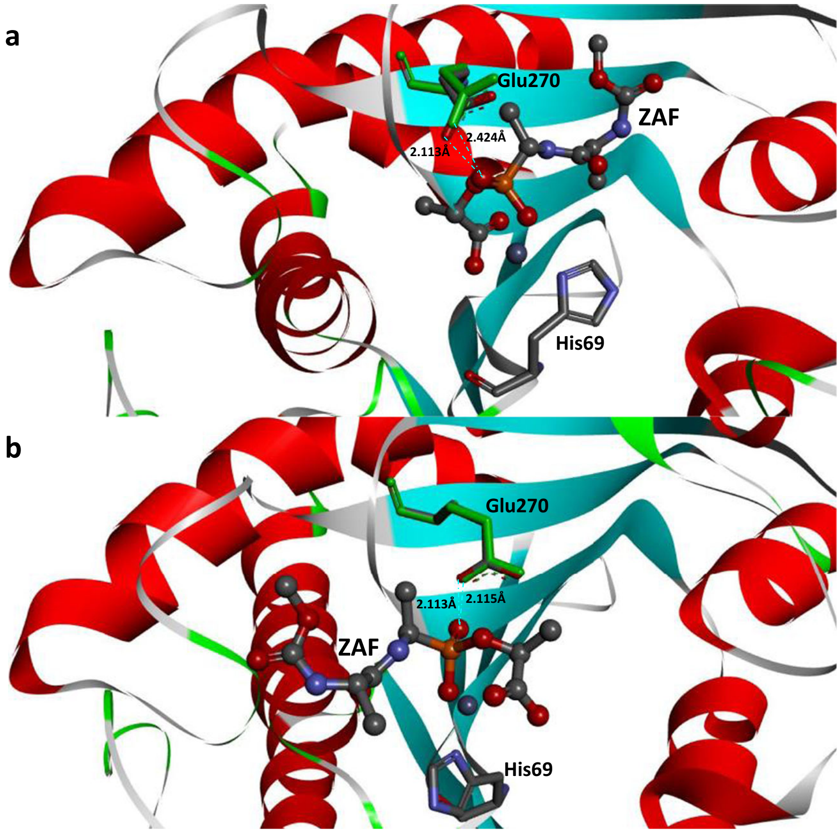


Fig. 1a–b Superposition of the matched and back-calculated geometries of the second catalytic residue in the main loop of scaffold 6CPA under different loop-closure resolutions: **a** 0.3 Å, **b** 0.05 Å. The matched transition state and catalytic residues are shown as *ball-and-stick models*, and the O, N, and C atoms are shown in *red*, *teal*, and *gray*, respectively. The back-calculated catalytic residue is shown as a *green stick model*. Distances are indicated by *dotted cyan lines*; values are in Å

angles, and torsion angles, such that atoms CB and CA of the second residue are superimposed on the corresponding atoms CB and CA at the fixed site. After the loop between the first two catalytic residues has been closed, the position and orientation of the transition state are determined, and loops from the transition state to the remaining catalytic residues are formed. Similar loop closure processes are performed to identify the conformations of the remaining residues and the parameters associated with the catalytic geometrical relationships between the transition state and the remaining catalytic residues. Mathematically, the optimization problem for loop closure in enzyme design can be formulated as follows:

$$\begin{aligned} \min \quad & f(x) = \sqrt{\frac{d_{\mathrm{CA}}^2 + d_{\mathrm{CB}}^2}{2}} \\ s.t. \quad & d^{\mathrm{L}} \le d \le d^{\mathrm{U}}, \\ & \theta^{\mathrm{L}} \le \theta \le \theta^{\mathrm{U}} \\ & \chi^{\mathrm{L}} \le \chi \le \chi^{\mathrm{U}} \end{aligned} , \qquad (1)$$

where the objective function is given by the root-mean squared deviation (RMSD) of atoms CA and CB of the moving residue, which is an implicit function of variables such as the bond length, bond angle, and torsion angle used to characterize the incumbent loop [i.e., $x=(d,\theta,\chi)$]. d_{CA} and d_{CB} are the distances between atoms CA and CB of the moving residue and the corresponding atoms at the fixed site. The lower and upper bounds (denoted by the superscript letters "L" and "U" in the equation) of the bond lengths and bond angles are given by the standard values and deviations of the bonds, and the lower and upper bounds of the torsion angles are always set to −180° and 180°. The initial structure of the loop can be constructed from the standard bond lengths and bond angles, while the initial values of the torsion angles are either 180°, 60°, or −60°. To solve optimization problem (1) quickly and with high accuracy, three pseudo-Newton-direction-based algorithms were developed that take advantage of the special mathematical structure of problem (1).

When the loop is long and has multiple torsions, it can be closed without the need to adjust the bond lengths and bond angles. Torsion angles are always free to rotate, and their values can be easily mapped into the interval [−180°, 180°] by virtue of the periodic characteristics of trigonometric functions with respect to torsions. Therefore, the bond lengths and bond angles are first fixed at their standard values, and problem (1) is transformed into an unconstrained optimization problem. An unconstrained optimization algorithm based on the quasi-Newton direction calculated by means of the BFGS formula [30] was developed to solve problem (1), as follows:

Algorithm I: unconstrained optimization algorithm

```
Step (0).  Starting point is $x^0$, convergence tolerance
           $\varepsilon = 10^{-5}$, inverse Hessian approximation
           $H^0 = I$, and count $k = 0$,
```

Step (1). If $\|\nabla f(x^k)\| < \varepsilon$ or RMSD < 0.05, stop. Otherwise continue.

Step (2). Compute the search direction $p^k = -H^k \nabla f(x^k)$.

Step (3). Compute the step length λ_k from a line search procedure as

$$\min_{\lambda} \ f\left(x^k + \lambda p^k\right). \tag{2}$$

Step (4). Set $x^{k+1} = x^k + \lambda_k p^k$, define $\Delta x^k = x^{k+1} - x^k$ and $\Delta G^k = \nabla f(x^{k+1}) - \nabla f(x^k)$, and compute H^{k+1} by means of the BFGS formula as

$$H^{k+1} = \left(I - \frac{\Delta x^k \left(\Delta G^k\right)^T}{\left(\Delta x^k\right)^T \Delta G^k}\right) H^k \left(I - \frac{\Delta G^k \left(\Delta x^k\right)^T}{\left(\Delta x^k\right)^T \Delta G^k}\right) + \frac{\Delta x^k \left(\Delta x^k\right)^T}{\left(\Delta x^k\right)^T \Delta G^k}. \tag{3}$$

where the superscript letter "T" indicates the transpose.

Step (5). Set $k = k + 1$ and go to step (1).

In algorithm I, the gradient vector $\nabla f(x^k)$ of the objective function $f(x)$ at x^k is computed using the finite difference method as the objective function of problem (1); that is, the RMSD of atoms CA and CB of the moving residues is an implicit function of the torsion angles:

$$\frac{\partial f}{\partial x_i}\left(x^k\right) = \frac{f\left(x^k + \varepsilon e_i\right) - f\left(x^k\right)}{\varepsilon}, \tag{4}$$

where ε is a small positive scalar and e_i is the i^{th} unit vector. In this work, the step length was computed by means of an inexact line search procedure based on the Armijo condition combined with quadratic backtracking [30, 31] as

Line search procedure

Step (0). Set $\lambda = 1$.

Step (1). Compute $f(x^k + \lambda p^k)$.

Step (2). If the Armijo condition $f(x^k + \lambda p^k) \le f(x^k) + \delta \lambda \nabla f(x^k)^T p^k$ holds, the step length is found, so stop. Otherwise continue.

Step (3). Let $\alpha = \max\{\eta, \lambda_q\}$, where

$$\lambda_q = \frac{\lambda \nabla f\left(x^k\right)^T p^k}{2\left(\lambda \nabla f\left(x^k\right)^T p^k + f\left(x^k\right) - f\left(x^k + \lambda p^k\right)\right)}. \tag{5}$$

Set $\lambda = \alpha \lambda$ and go to step (1).

The parameters δ and η in the above procedure are set to 0.0001 and 0.1. Parameter η is introduced to avoid obtaining a step length that is too small; it adds robustness and reliability to quasi-Newton methods, particularly if the starting point is a poor choice. Another issue is that the objective function in problem (1) is nonlinear and nonconvex, so the quadratic interpolation of the step length approximation given by Eq. (5) may produce incorrect results. To account for this, a protection measure can be used to force $\lambda_q = 0.1$ if $f(x^k + \lambda p^k) < f(x^k) + \lambda \nabla f(x^k)^T p^k$, which avoids entry to a dead loop if the objective function at the incumbent point is concave.

When the loop is short, the bond lengths and bond angles must be adjusted to close the loop with high accuracy. In this case, loop closure problem (1) can be reformulated as a general bound-constrained optimization problem in order to present the necessary conditions for optimal solutions explicitly as

$$\begin{array}{ll} \min & f(x) \\ s.t. & l_i \le x_i \le u_i, \quad i = 1, 2, \ldots, n \end{array}, \tag{6}$$

where l_i and u_i represent the lower and upper bounds on the variables x_i. When the variable, such as the torsion angle, has no bounds, the lower and upper bounds are set to ∓ infinity. The bound-constrained optimization problem is a special case of the constrained optimization problem [32], and the conditions required for optimal solutions can be derived from the Karush–Kuhn–Tucker conditions for problem (6) and simplified as follows:

$$\begin{array}{l} \nabla f_i\left(x^*\right) \ge 0, \quad \text{when } x_i^* = l_i \\ \nabla f_i\left(x^*\right) = 0, \quad \text{when } l_i < x_i^* < u_i, \\ \nabla f_i\left(x^*\right) \le 0, \quad \text{when } x_i^* = u_i \end{array} \tag{7}$$

where x^* is one of the optimal solutions of problem (6). The projected gradient was defined as

$$\nabla \overline{f}(x) = \begin{cases} \nabla \overline{f}_i(x) = \min\{0, \nabla f_i(x)\} & \text{if } x_i = l_i \\ \nabla \overline{f}_i(x) = \min\{0, \nabla f_i(x)\} & \text{if } l_i < x_i < u_i \\ \nabla \overline{f}_i(x) = \max\{0, \nabla f_i(x)\} & \text{if } x_i = u_i \end{cases}. \tag{8}$$

A projected quasi-Newton-direction-based active-set algorithm was developed in this work to solve problem (6) as follows:

Algorithm II: bound-constrained optimization algorithm

Step (0). Starting point is x^0, convergence tolerance $\varepsilon = 10^{-5}$, inverse Hessian approximation $H^0 = I$, and count $k = 0$,

Step (1). If the projected gradient at the incumbent point satisfies $\left\|\nabla \overline{f}\left(x^k\right)\right\| < \varepsilon$ or RMSD < 0.05, stop. Otherwise continue.

Step (2) Compute the quasi-Newton direction $\overline{p}^k = -H^k \nabla f\left(x^k\right)$, and the search direction can be obtained by means of projection as

$$p^k = \begin{cases} p_i^k = 0, & \text{if } x_i^k = l_i \text{ and } \overline{p}_i^k < 0, \\ p_i^k = 0, & \text{if } x_i^k = u_i \text{ and } \overline{p}_i^k > 0, \\ p_i^k = \overline{p}_i^k, & \text{otherwise.} \end{cases} \tag{9}$$

Step (3). Compute the step length λ_k using the same line search procedure, i.e., $\min_{\lambda} f\left(x^k + \lambda p^k\right)$, as that in Algorithm I, but compute the initial step length via

$$\lambda^0 = \min\left\{1.0,\ \min_{p_i^k<0}\left\{\frac{l_i - x_i^k}{p_i^k}\right\},\ \min_{p_i^k>0}\left\{\frac{u_i - x_i^k}{p_i^k}\right\}\right\}. \tag{10}$$

Step (4). Set $x^{k+1} = x^k + \lambda_k p^k$, define $\Delta x^k = x^{k+1} - x^k$ and $\Delta G^k = \nabla f(x^{k+1}) - \nabla f(x^k)$, and compute H^{k+1} by means of the BFGS formula, i.e., Eq. 3.

Step (5). Set $k = k+1$ and go to step (1).

Similar to Algorithm I, the gradient vector $\nabla f(x^k)$ is computed by means of the finite difference method described by Eq. 4. In the bound-constrained optimization algorithm, the bond lengths and bond angles are adjusted simultaneously with the torsion angles. Consequently, in the optimal solutions to problem (1) obtained by the bound-constrained algorithm, the bond lengths and bond angles prefer to lie at the boundaries and significantly deviate from the standard values. To overcome this issue, a combined algorithm that integrates the unconstrained optimization algorithm with the bound-constrained optimization algorithm was developed to solve problem (1):

Algorithm III: combined optimization algorithm

```
Step (0). Starting point is $x^0$, convergence tolerance $\varepsilon = 10^{-5}$, maximum number of iterations is predefined, and count $k = 0$. The initial bond lengths and bond angles are set to standard values.
Step (1). If count $k$ is greater than the maximum number of iterations and RMSD > 0.05, stop; loop closure has failed. Otherwise continue.
Step (2). Run the unconstrained optimization algorithm (Algorithm I). If RMSD < 0.05, stop. Otherwise continue.
Step (3). Fix the torsion angles at the optimal solutions obtained in step (2) and continue.
Step (4). Run the bound-constrained optimization algorithm (Algorithm II) with fixed torsion angles. If RMSD < 0.05, stop. Otherwise continue.
Step (5). Fix the bond lengths and bond angles at the optimal solutions obtained in step (3) and continue.
Step (6). Set $k = k+1$ and go to step (1).
```

In the inner loops of the combined algorithm, the maximum number of iterations of the unconstrained optimization algorithm and the bound-constrained optimization algorithm are set to 30 and 50, respectively. The maximum number of iterations of the combined algorithm in the outer loop is set to 5, and the combined algorithm is evaluated by the total number of iterations, which is the sum of the iterations of the unconstrained optimization algorithm and the iterations of the bound-constrained optimization algorithm in the inner loops of the combined algorithm.

Results and discussion

Impact of loop-closure accuracy on the catalytic geometrical constraints

In the matching algorithm of ProdaMatch, the sites and geometries of the catalytic residues are determined by the solutions to loop-closure problems, where the main loop consists of the first two catalytic residues bridged by the TS, while each of the side loops is composed of the atoms of one remaining catalytic residue and the corresponding atoms of the TS. The coordinates of the atoms CA and CB of all but the first catalytic residue are obtained from the loop-closure calculation, and the calculated CA and CB atoms may deviate from the fixed atoms in the crystal structure depending on the computing tolerance set in the loop-closure algorithm. However, the geometries of the catalytic residues will ultimately grow from the CA and CB atoms in the crystal structure, and the inaccuracy of the calculation of the loop-closure algorithm (such as CCD, as used in the original ProdaMatch) may lead to the violation of the catalytic geometrical constraints between the TS and the catalytic residues. To explicitly illustrate this drawback, we used the 6CPA case study from the benchmark test set compiled by Zanghellini et al. [14] to validate the matching algorithms employed in enzyme design. The PDB (Protein Data Bank) code 6CPA represents a bovine carboxypeptidase A type of hydrolase with four catalytic residues in its active pocket—His69, Glu72, His196, and Glu270—which were used to match with the transition-state analog from carboxypeptidase-A-catalyzed hydrolysis, ZAF. The catalytic residues His69 and Glu270 together with the TS comprise the main loop in the matching process, and the geometries of His69 and Glu270 as well as the position and orientation of the TS are determined by the solution to the loop-closure problem. During the loop-closure process, the torsion angles of the catalytic residues and those involved in the catalytic geometrical relationships between the catalytic residues and the TS are free to rotate, and the standard bond lengths and bond angles are taken directly from the crystal structure 6CPA. The allowed ranges of the bond lengths and bond angles correspond to deviations from the standard values of ±0.1 Å and ±10°, respectively. In the main loop, the coordinates of the atoms CB and CA from the second catalytic residue are determined by the loop-closure problem, so the catalytic geometrical constraints between the TS and the second catalytic residue are directly affected by the loop-closure tolerance. The standard values of these constraints, i.e., the bond angles between atom P22 of ZAF, atom O20 of ZAF, and atom OE2 of Glu270 and between atom O20 of ZAF, atom OE2 of Glu270, and atom CD of Glu270, as well as the bond length between atom O20 of ZAF and atom OE2 of Glu270 and the matched results obtained from ProdaMatch at different loop-closure tolerances are given in Table 1. The true values of the catalytic geometrical constraints were obtained based on the coordinates of Glu270, which were back-calculated starting from the crystal backbone atoms N, CA, and CB, but using the matched torsion angles. Here, the three matched torsion angles of Glu270 were calculated via the matched coordinates of atoms OE1, CD, CG, CB, and CA and the coordinates of atom N in the crystal structure. When the loop-closure tolerance was set to 0.3 Å, the catalytic geometrical constraints between the TS and Glu270 were violated, though the matched values were in the feasible range. Specifically, the true value of the bond angle between atom P22 of ZAF, atom O20 of ZAF, and atom OE2 of Glu270 was less than the lower bound of its feasible range, and the true value of the bond length between atom O20 of ZAF and atom OE2 of Glu270 was greater than the upper bound of its feasible range. The matched and back-calculated geometries of Glu270 are shown in Fig. 1a, where the bond lengths between atom O20 of ZAF and atom OE2 of Glu270 are presented. However, such violations disappeared when a stricter loop-closure tolerance was used. Based on the results shown in Table 1 and Fig. 1b, the matched and back-calculated results for the catalytic geometrical constraints between TS and Glu270 all lie in the feasible ranges when the loop-closure tolerance is set to 0.05 Å. Moreover, the matched and back-calculated geometries of Glu270 almost overlap with each other. However, the number of iterations of the original loop-closure algorithm CCD will inevitably increase with higher tolerance resolution. The relationship between the number of iterations of the CCD algorithm and the loop-closure tolerance is shown in Fig. 2, where the data were collected from 1000 initial loops under different loop-closure tolerances. When the loop-closure tolerance is set to 0.05 Å, >500 iterations are needed by the CCD algorithm to close the 188 initial loops. This is obviously not acceptable because thousands of loops need to be closed in the matching process. From an algorithmic perspective, CCD is a steepest descent algorithm, and the zigzagging phenomenon always appears when it approaches the minimum point as the step length is determined by the accurate line search method. As a result, CCD always requires a greater number of iterations with a high convergence tolerance. To get achieve a good balance between speed and accuracy, a quasi-Newton-direction-based novel algorithm was developed to close the loops quickly under high convergence tolerance.

Loop-closure results using the combined algorithm

The loop-closure performance of the combined algorithm was assessed using the benchmark test set for the recapitulation of native active sites compiled by Zanghellini et al. [14], which was also used in our earlier work [17] to validate the original ProdaMatch. Ten crystal structures in the test set as well as the corresponding catalytic residues in their active sites are shown in Table 2, and all loops to be closed in the matching algorithm for these ten scaffolds are also presented in Table 2. Assuming that we have n catalytic residues to match with a particular scaffold, we need to close one main loop and $(n-2)$ side loops for this scaffold based on the definition of the loop-closure problem in the matching process. When the number of

Table 1 Matched and back-calculated results obtained with different loop-closure resolutions for the catalytic geometrical constraints between ZAF and Glu270 in scaffold 6CPA

Catalytic geometrical constraints [a]						Standard values[b] (deviation)	0.3 Å		0.05 Å	
Site pair	Type	Atom1	Atom2	Atom3	Atom4		Matched	Back-cal [c]	Matched	Back-cal [c]
ZAF–Glu270	Angle	P22	O20	#OE2	–	136.84 (±10°)	126.84°	125.65°	139.38°	139.57°
	Angle	O20	#OE2	#CD	–	123.45 (±10°)	113.45°	123.26°	117.48°	121.89°
	Length	O20	#OE2	–	–	2.21 (±0.1 Å)	2.11 Å	2.42 Å	2.11 Å	2.12 Å
	Torsion angle	CH3	P22	O20	#OE2	80.78° (free)	56.67°	41.84°	53.19°	57.26°
	Torsion angle	P22	O20	#OE2	#CD	60.16° (free)	−53.00°	−28.93°	77.74°	72.60°
	Torsion angle	O20	#OE2	#CD	#CG	−155.93° (free)	127.36°	111.41°	−156.51°	−153.30°

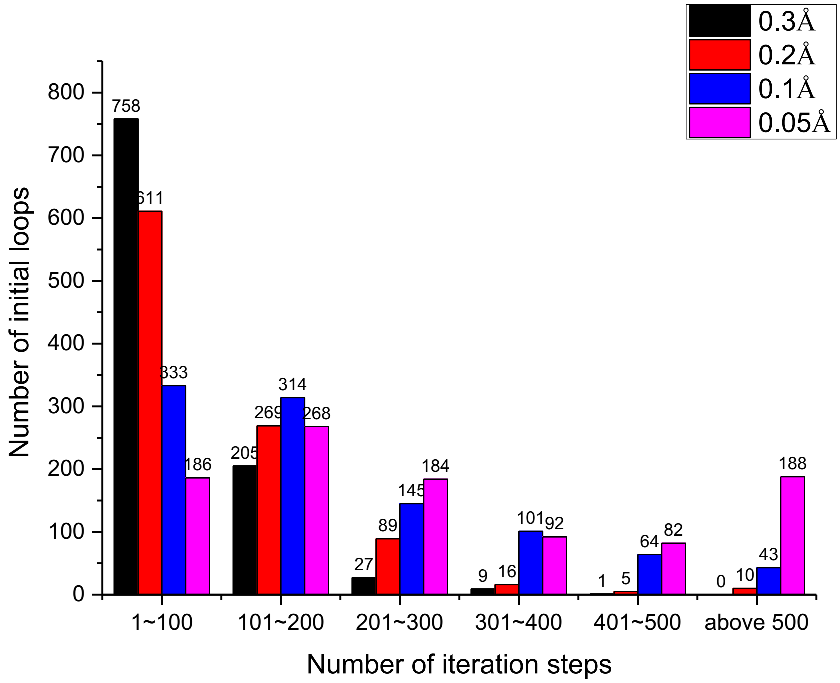
Lengths are given in Å. Angles and torsion angles are given in degrees

[a] Atoms on the latter residue in a site pair are prefixed with #

[b] Standard values are derived directly from the crystal structures

[c] Columns labeled "Back-cal" represent catalytic geometrical parameters which are back-calculated by using the matched geometrical values to grow the side chain of Glu270 from its backbone

Fig. 2 Histograms for number of iteration steps and number of initial loops closed by the CCD algorithm for the main loop of scaffold 6CPA under different resolutions



catalytic residues is greater than two, selection of the two catalytic residues for constructing the main loop for each scaffold proceeds according to three criteria: (i) the two catalytic residues should simultaneously have catalytic geometrical relationships with the TS; (ii) the catalytic residues are selected preferentially if they accurately position the TS based on the catalytic geometrical relationships between them; and (iii) the catalytic residues with the fewest torsion angles are preferred. In total, we have 21 loops to close. The RMSDs of the TS structures calculated by the loop-closure processes for main loops may hinder the proper evaluation of the algorithmic performance of the novel algorithm for closing side loops. Therefore, in this work, all side loops grew from the functional atoms of the TS or preceding catalytic residues in the crystal structure instead of their calculated structures. For each loop in the test set, multiple initial loops were generated depending on the number of torsion angles in this loop. The bond angles and bond lengths of each loop were set to their standard values. However, each torsion angle of the loop, which came from either the catalytic geometrical relationship between the catalytic residue and the TS or from the geometries of the catalytic residues, was assigned one of three discrete values: 180°, +60°, or −60°. Therefore, in total, 3^k initial loops were generated for each loop, where k is the number of torsion angles in the loop. To limit the total number of initial loops for very long loops to a maximum of eight torsion angles, all but the first eight torsion angles were set to 180°.

The loop-closure results obtained upon using the combined algorithm for the 21 loops in the test set are shown in Table 2. The loop was considered closed when the RMSD of the CB and CA atoms of the moving and fixed catalytic residues was less than 0.05 Å. Of the 64,827 initial loops generated from the main or side loops, 99.51 % of them closed to within an RMSD of 0.05 Å in fewer than 400 steps. Among the closed initial loops, 72.13 % of them were only closed by running the unconstrained optimization algorithm, indicating that the bond lengths and bond angles retained their standard values during the loop-closure processes to maintain the perfect geometries. The proportion of the side loops that closed, 91.33 %, is lower than that of the main loops, 99.99 %. Moreover, among the closed side loops, 51.89 % were only closed by running the unconstrained optimization algorithm, which is much lower than the 73.22 % of the main loops which were closed. The poor loop-closure performance of the side loops was caused by the fact that the loop lengths of the side loops are always much shorter than those of the main loops (here, the loop length is defined by the number of torsions, the number of angles, and the number of lengths in a loop). As shown in Table 2, in the "Loop length" column, the shortest main loop had seven torsion angles, but the longest side loop had only six freely rotatable torsion angles. To close the side loops within the convergence tolerance, the bounded bond length and bond angle variables had to be adjusted by running the bound-constrained optimization algorithm.

The maximum number of iterations of the combined algorithm was limited to 400, as discussed in the "Materials and methods" section. Histograms for the number of iteration steps required to close the loops in the test set are shown in Fig. 3 for the scaffolds 1NEY and 3VGC, whereas those for other scaffolds are presented in Fig. S1 in the "Electronic supplementary material" (ESM). The large majority of the initial loops closed within 100 steps, which was almost tenfold

Table 2 Loop-closure results for ten scaffolds obtained by novel optimization algorithms

PDB	Catalytic residues and TS	Loop name	Loop length [a]	Number of loops	Number of closed loops [b]	Number of loops closed by BFGS [c]	Number of loops closed by CCD	Number of TSs with RMSD < 2.0 Å	Min. RMSD of TSs
1C2T	Asn106, Asp144, His108, NHS	Asn106_NHS_Asp144	(9,4,2)	6561	6561 (100 %)	4838 (73.74 %)	6534 (99.59 %)	11	0.99
		NHS_His108	(4,2,1)	81	50 (61.73 %)	42 (51.85 %)	66 (81.48 %)		
1DQX	Asp91, Asp273, Lys93, Lys59, BMP	Asp91_BMP_Asp273	(9,4,2)	6561	6561 (100 %)	5816 (88.65 %)	6558 (99.95 %)	170	0.54
		Asp91_Lys93	(6,2,1)	729	702 (96.30 %)	424 (58.16 %)	648 (88.89 %)		
		Asp91_Lys59	(6,2,1)	729	655 (89.85 %)	0.00 %	709 (97.26 %)		
1H2J	Glu136, Glu225, DCB	Glu136_DCB_Glu225	(11,4,2)	6561	6561 (100 %)	5751 (87.65 %)	6561 (100 %)	2	1.88
1JCL	Lys168, Asp103, Lys202, HPD	Lys168_HPD_Asp103	(11,4,2)	6561	6561 (100 %)	5391 (82.17 %)	6561 (100 %)	9	0.72
		Asp103_Lys202	(6,2,1)	729	725 (99.45 %)	486 (66.67 %)	711 (97.53 %)		
1NEY	Glu164, His94, Lys11, 13P	Glu164_13P_His94	(10,4,2)	6561	6560 (99.98 %)	3674 (56.00 %)	4908 (74.81 %)	187	0.35
		13P_Lys11	(6,2,1)	729	728 (99.86 %)	486 (66.67 %)	729 (100 %)		
1OEX	Asp217, Asp35, LOV	Asp217_LOV_Asp35	(9,4,2)	6561	6561 (100 %)	4445 (67.75 %)	6507 (99.18 %)	113	0.58
1P6O	Cys89, Glu62, His60, Cys92, HPY	Cys89_HPY_Glu62	(9,4,2)	6561	6561 (100 %)	5767 (87.90 %)	6561 (100 %)	83	0.64
		HPY_His60	(4,2,1)	81	61 (75.31 %)	32 (39.51 %)	63 (77.78 %)		
		HPY_Cys92	(3,2,1)	27	6 (22.22 %)	5 (18.52 %)	27 (100 %)		
3VGC	Ser175, His42, Asp87, SRB	Ser175_SRB_His42	(7,4,2)	2187	2186 (99.95 %)	1622 (74.17 %)	2001 (91.50 %)	22	0.35
		His42_Asp87	(4,2,1)	81	59 (72.84 %)	49 (60.49 %)	81 (100 %)		
4FUA	His155, His94, His92, PGH	His155_PGH_His94	(9,4,2)	6561	6558 (99.95 %)	3754 (57.22 %)	5445 (82.99 %)	309	0.35
		PGH_His92	(4,2,1)	81	63 (77.78 %)	25 (30.86 %)	45 (55.56 %)		
6CPA	His69, Glu270, His196, Glu72, ZAF	His69_ZAF_Glu270	(10,4,2)	6561	6560 (99.98 %)	3776 (57.55 %)	4773 (72.75 %)	1163	0.61
		ZAF_His196	(4,2,1)	81	26 (32.10 %)	22 (27.16 %)	69 (85.19 %)		
		ZAF_Glu72	(5,2,1)	243	205 (84.36 %)	131 (53.91 %)	234 (96.30 %)		
In total		all Loops	(7.0,3.0,1.5)	64,827	64,510 (99.51 %)	46,536 (71.78 %)	59,791 (92.23 %)		
		main Loops	(9.4,4.0,2.0)	61,236	61,230 (99.99 %)	44,834 (73.22 %)	56,409 (92.12 %)		
		side Loops	(4.7,2.0,1.0)	3591	3280 (91.34 %)	1702 (47.40 %)	3382 (94.18 %)		

[a] The numbers in parentheses in the column "Loop length" refer to the number of torsion angles, the number of angles, and the number of lengths, which were determined by the loop-closure algorithm

[b] The values in parentheses are the ratio of the number of loops closed by the combined optimization algorithm to the number of initial loops

[c] The values in parentheses are the ratio of the number of loops closed by the unconstrained optimization algorithm to the number of initial loops

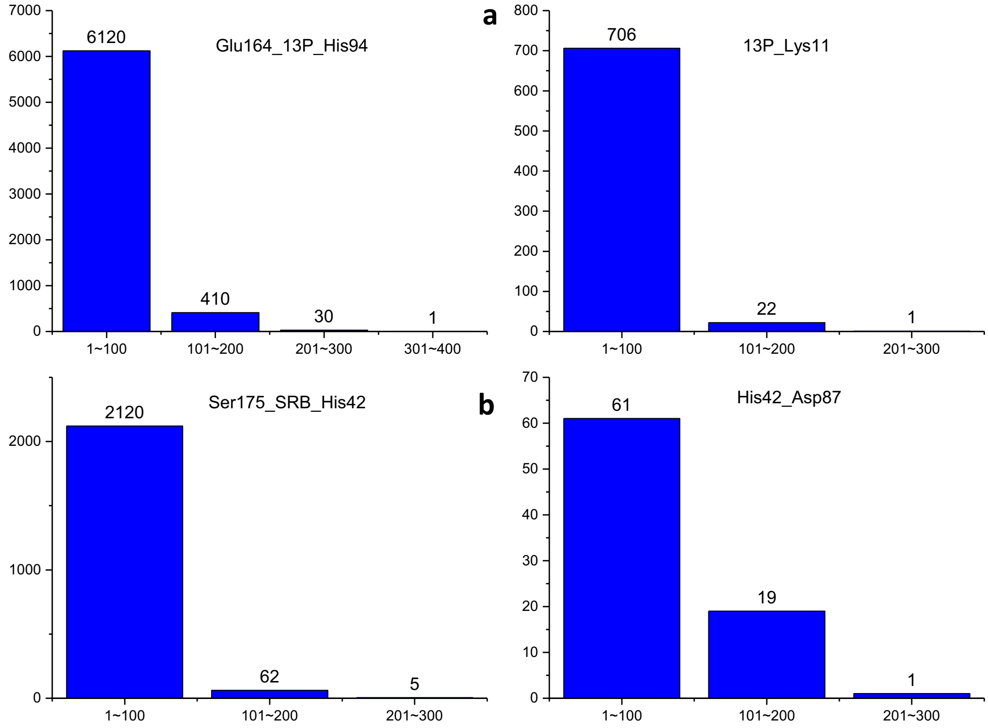


Fig. 3a–b Histograms of number of iteration steps and number of initial loops closed by the novel loop-closure algorithm to within an RMSD of 0.05 Å for scaffolds **a** 1NEY and **b** 3VGC. In all histograms, the abscissa shows the number of iteration steps while the ordinate shows the number of initial loops

the speed of the CCD algorithm using the same convergence tolerance. Furthermore, the proportion of the main loops that were closed, 92.12 %, when using CCD with the same tolerance and iteration limit was much lower than that of the combined algorithm (99.99 %), although the proportion of the side loops that were closed using CCD (94.18 %) is moderately higher than that achieved using the combined algorithm (91.34 %). For all initial loops, the number of steps required to close side loops was higher than the number required to close main loops. The side loops always had fewer torsion angles than the main loops, and the unconstrained optimization algorithm alone could not close the loops to within an RMSD of 0.05 Å. Therefore, the bound-constrained optimization algorithm was always invoked to further reduce the RMSD. This was a slower process than the unconstrained optimization algorithm, as its step length was restricted by the upper and lower bounds on bond lengths and bond angles. However, the large majority of the initial side loops were closed within 200 steps to within an RMSD of 0.05 Å. This implies that the combined optimization algorithm can close loops very quickly, and its accuracy can be evaluated by comparing the RMSDs of the TS with those in the crystal structure. As shown in Table 2, many of the closed main loops for all of the scaffolds position the TS to an RMSD of <2.0 Å. The best RMSDs of the TS for all of the scaffolds except for 1H2J were <1.0 Å. The matched and crystal geometries of scaffolds 1NEY and 3VGC are shown in Fig. 4, and those of other scaffolds are presented in Fig. S2 of the ESM. The matched TS geometries in scaffolds 1NEY and 3VGC overlap almost completely with those in their respective crystal structures and are supported by the RMSD values given in Table 2 (i.e., 0.353 for 1NEY and 0.349 for 3VGC). This indicates that the novel loop-closure algorithm can anchor the active site very accurately, but it still needs to be coupled with an energy-scoring function that can be used to identify either the closed main loop with the smallest RMSD of the TS or the loops with the smallest RMSDs.

Recapitulation of native active sites

The novel loop-closure algorithm described herein was designed to close loops in the matching process for catalytic residue site selection during enzyme design. It was imbedded into our matching algorithm, ProdaMatch, and implemented in the program PRODA, the PROtein Design Algorithmic package [17, 33, 34]. The native active-site recapitulation results obtained by the revised ProdaMatch for ten scaffolds in the benchmark set compiled by Zanghellini et al. [14] are presented in Table 3. They were obtained by running ProdaMatch on a single 2.1-GHz central processing unit (CPU) from a computer cluster with 64 cores that shared 128 GB of random-access memory. For each scaffold, heavy atoms of the catalytic residues that were found to lie within 5 Å of the TS in the crystal structure were selected as candidate sites to anchor the catalytic residues. The number of candidate sites for each scaffold is given in Table 3. The number of combinations of catalytic residues attached to the sites is large, which would greatly increase the time spent searching if all of the initial loops generated for each loop were closed. Thus, in our matching algorithm, only some of the initial loops

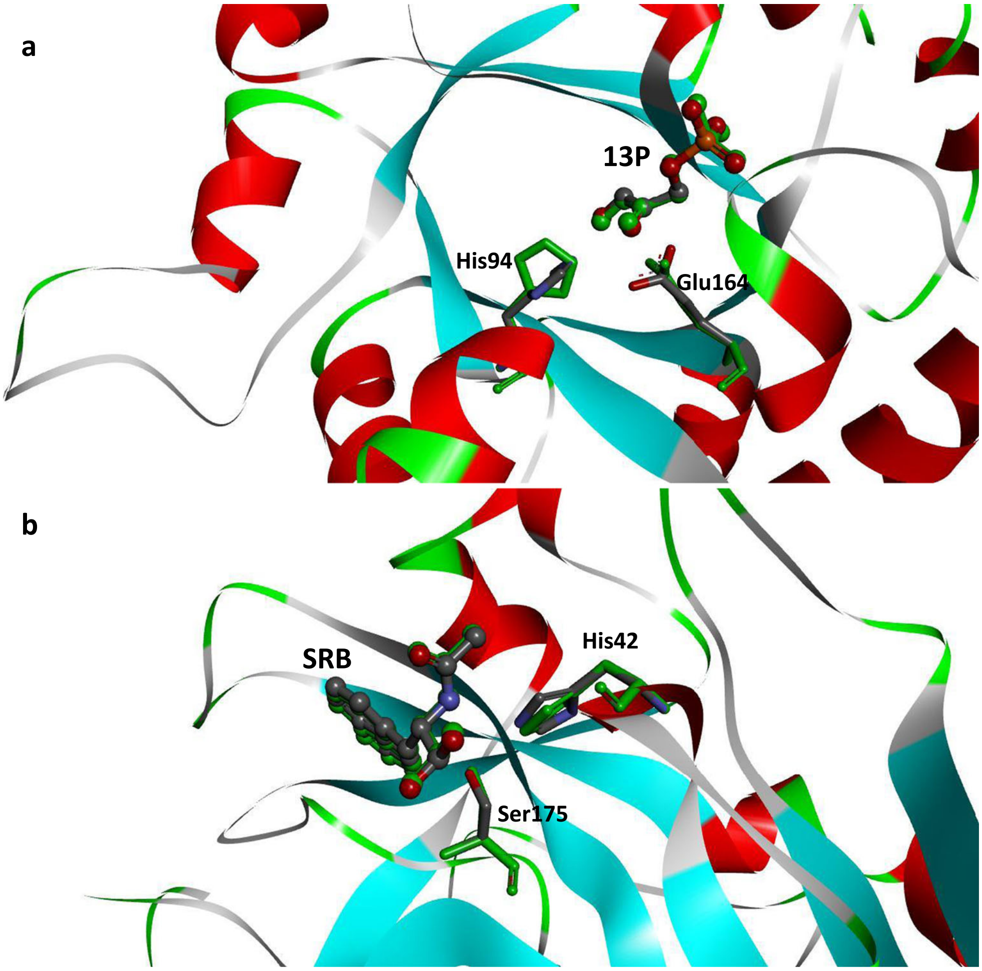


Fig. 4a–b Superposition of the matched and crystal geometries for main loops of the scaffolds **a** 1NEY and **b** 3VGC. The transition state (TS) and catalytic residues in the crystal structures are shown as *ball-and-stick models*, and the O, N, and C atoms are shown in *red*, *teal*, and *gray*, respectively. The matched transition state and catalytic residues are shown as *cyan stick models*

for each loop are closed. The initial loops generated for each loop are ranked by an energy score which is calculated as the sum of (i) the intrinsic collision energies of the catalytic residues; (ii) the interaction energy between the catalytic residues and TS; and (iii) the interaction energy between the scaffold template and the catalytic residues as well as the

Table 3 Recapitulation of native active sites using the revised ProdaMatch for ten scaffolds

PDB	Catalytic residues	Number of sites [a]	Number of matches	Rank[b]	Number of matches using CCD (rank)[c]	RMSD of TS (Å) [d]	CPU time
1C2T	Asn106, His108, Asp144	27	28	20	456 (162)	7.56	20 h 50 m
1DQX	Asp91, Asp273, Lys93, Lys59	29	16,164	37	28,826 (1034)	1.79	20 h 59 m
1H2J	Glu225, Glu136	24	190	39	194 (60)	3.94	9 h 49 m
1JCL	Lys168, Asp103, Lys202	23	709	56	432 (234)	3.77	24 h 20 m
1NEY	Glu164, His94, Lys11	20	222	2	245 (−)	1.73	3 h 56 m
1OEX	Asp35, Asp217	21	35	4	71 (27)	6.31	25 m
1P6O	Cys89, Cys92, His60, Glu62	18	28	10	1719 (392)	1.25	1 h 14 m
3VGC	Ser175, His42, Asp87	34	26	7	8 (−)	1.75	3 h 58 m
4FUA	His155, His94, His92	19	83	1	98 (−)	2.96	3 h 53 m
6CPA	His69, Glu270, His196, Glu72	23	156	1	266 (1)	1.47	24 h 7 m

[a] The column “Number of sites” shows the number of candidate sites for anchoring the catalytic residues

[b] The column “Rank' shows the rank of the native match among all identified matches

[c] A dash in parentheses indicates that the native matches were not identified

[d] The column “RMSD of TS” shows the RMS deviation of the matched TS from that in the crystal structure

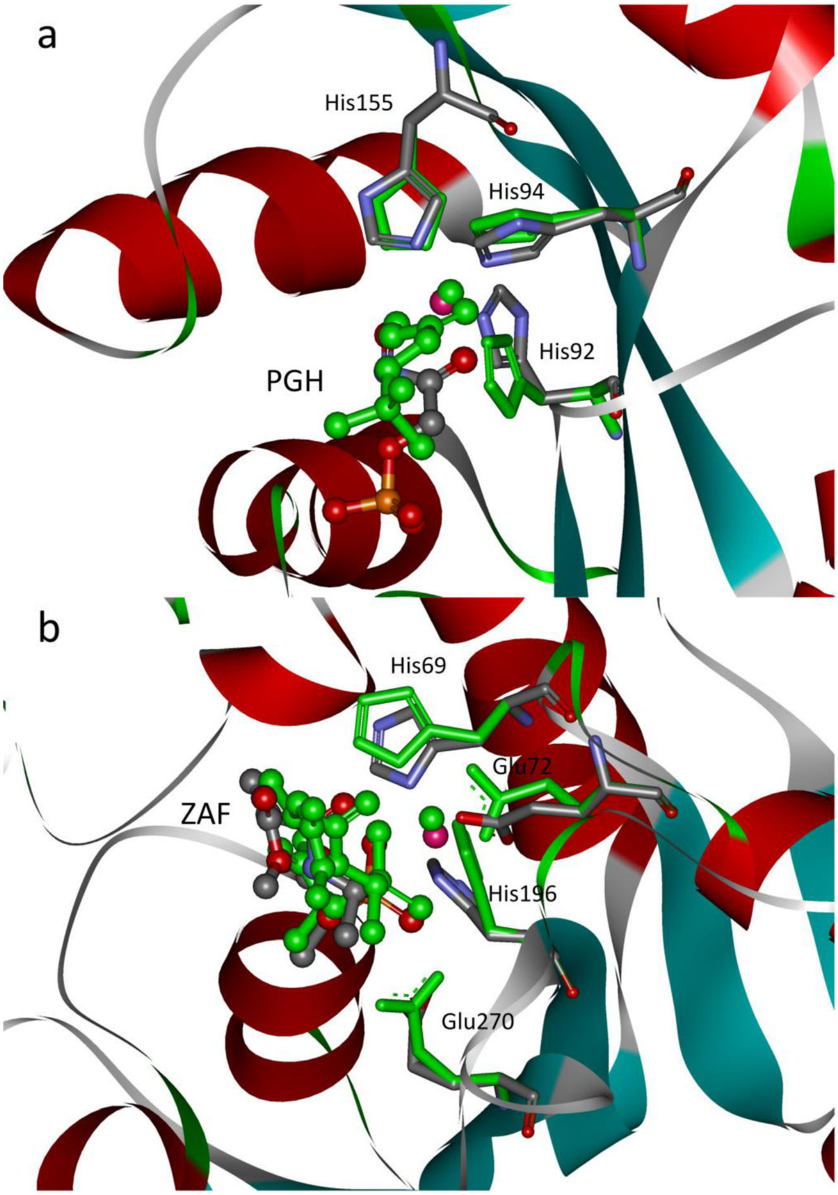


Fig. 5a–b Superposition of native and predicted active sites: **a** scaffold 4FUA; **b** scaffold 6CPA. The transition state (TS) and catalytic residues in the crystal structures are shown as *ball-and-stick models*, and the O, N, and C atoms are shown in *red*, *teal*, and *gray*, respectively. The matched transition state and catalytic residues are shown as *cyan stick models*

TS. The interaction energy between atoms is calculated by a linear repulsive term described by Equation 2 in our previous work [35]. Up to 1000 initial loops for each loop are closed using the combined algorithm. The purpose of the above screening process is to eliminate many of the initial loops which intrinsically collide with the scaffold template.

Table 3 shows that all native matches for the ten scaffolds were identified by the revised ProdaMatch, which utilizes the novel loop-closure algorithm to close loops, although the loop-closure tolerance was decreased sixfold, from 0.3 to 0.05 Å. The native matches rank in the top ten of all of the identified matches for the six scaffolds. The number of matches and the ranks of the native matches obtained by the ProdaMatch algorithm using CCD (as shown in Table 3) indicate that the matching results obtained using CCD with the same tolerance and iteration limit were very poor, and the native matches for scaffolds 1NEY, 3VGC, and 4FUA were not identified. The TS positions and the side-chain geometries of the catalytic residues in the calculated native matches were well constructed, and the RMSDs of the identified TSs in five scaffolds for native matches were less than 2.0 Å. The native active-site recapitulations for two scaffolds, 1NEY and 6CPA, which both ranked as the top native match among all identified matches and had low RMSDs of the TSs, are shown in Fig. 5. ProdaMatch, by virtue of the novel loop-closure algorithm, can position the TS in five scaffolds. However, in some scaffolds, such as 1C2T and 1OEX (shown in Fig. S3 of the ESM), the position of the TS deviates significantly from that in the crystal structure, although the catalytic geometrical relationships between the TS and the catalytic residues are all satisfied. In Table 2, the best RMSDs of the TSs identified by the closed main loops for all scaffolds are less than 2.0 Å, and nine of them are less than 1.0 Å. These discrepancies occur because only 1000 initial loops for each main loop are selected at most by the proposed scoring function for closure in the matching processes, but all of them are closed for the loop-closure test, as shown in Table 2. Clearly, the initial loops that resulted in the best RMSDs of the TS in the loop-closure test were missed by our scoring function in the native active-site recapitulation test. As accurately establishing the position of the TS is critical for enzyme catalysis in de novo enzyme design, more effective scoring functions for selecting initial loops as the main loops are still required. The CPU time spent on the test cases by the revised ProdaMatch is shown in Table 3; all matching processes for all scaffolds were completed in approximately one day. Parallel computation is easy to implement, and a previously expensive computer cluster with several hundred CPU cores is now relatively inexpensive. As such, ProdaMatch—by virtue of the novel loop-closure algorithm—can be used for scaffold selection in enzyme design projects where several thousand PDB scaffolds are included in a scaffold library.

Conclusions

Herein, a Newton-direction-based novel loop-closure algorithm was developed to close loops in the matching of catalytic residues with a scaffold during de novo enzyme design. It was shown to be faster than the CCD-based loop-closure algorithm when using higher convergence tolerances, and to eliminate the potential violation of catalytic geometrical constraints between the TS and catalytic residues caused by the coarse loop-closure tolerance set in the original ProdaMatch algorithm. Among the 64,824 initial loops derived from 21 loops of the enzyme design test set, 99.51 % were closed by the novel loop-closure algorithm to within an RMSD of 0.05 Å, while the large majority of the initial loops were closed within 100 iteration steps, which is tenfold faster than achieved with the CCD algorithm using the same convergence tolerance. In the native active-site recapitulation test, the revised ProdaMatch identified all native matches for ten scaffolds, although the loop-closure tolerance was strengthened from 0.3 to 0.05 Å. Although the native match ranked high among all identified matches, further development of a scoring function for effectively selecting the initial loops should improve the accuracy of the position of the TS during native active-site recapitulation. The revised ProdaMatch algorithm could potentially be used for scaffold selection because of its high accuracy and fast speed.

Acknowledgments This work was financially supported by the National Natural Science Foundation of China (grant numbers: 21276136, 21476123), and the National High Technology Research and Development (863) Program of China (grant number: 2012AA021204).

References

1. Walsh C (2001) Enabling the chemistry of life. Nature 409: 226–231
2. Wolfenden R, Snider MJ (2001) The depth of chemical time and the power of enzyme as catalyst. Acc Chem Res 34:938–945
3. Schmid A, Dordick JS, Hauer B, Kiener A, Wubbolts M, Witholt B (2001) Industrial biocatalysis today and tomorrow. Nature 409: 258–268
4. Arnold FH (2001) Combinatorial and computational challenges for biocatalyst design. Nature 409:253–257
5. Toscano MD, Woycechowsky KJ, Hilvert D (2007) Minimalist active-site redesign: teaching old enzymes new tricks. Angew Chem Int Ed 46:3212–3236
6. Bornscheuer U, Huisman G, Kazlauskas R, Lutz S, Moore J, Robins K (2012) Engineering the third wave of biocatalysis. Nature 485:185–194
7. Kiss G, Çelebi-Ölçüm N, Moretti R, Baker D, Houk KN (2013) Computational enzyme design. Angew Chem Int Ed 52:5700–5725
8. Khoury GA, Smadbeck J, Kieslich CA, Floudas CA (2014) Protein folding and de novo protein design for biotechnological applications. Trends Biotechnol 32:99–109
9. Hilvert D (2013) Design of protein catalysts. Annu Rev Biochem 82:447–470
10. Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D (2011) De novo enzyme design using Rosetta3. PLoS One 6:e19230
11. Tantillo DJ, Chen J, Houk KN (1998) Theozymes and compuzymes: theoretical models for biological catalysis. Curr Opin Chem Biol 2:743–750
12. Hellinga HW, Richards FM (1991) Construction of new ligand binding sites in proteins of known structure: I. Computer-aided modeling of sites with pre-defined geometry. J Mol Biol 222: 763–785
13. Lassila JK, Privett HK, Allen BD, Mayo SL (2006) Combinatorial methods for small-molecule placement in computational enzyme design. Proc Natl Acad Sci USA 103:16710–16715
14. Zanghellini A, Jiang L, Wollacott AM, Cheng G, Meiler J, Althoff EA, Röthlisberger D, Baker D (2006) New algorithms and an in silico benchmark for computational enzyme design. Protein Sci 15: 2785–2794
15. Fazelinia H, Cirino PC, Maranas CD (2009) Optgraft: a computational procedure for transferring a binding site onto an existing protein scaffold. Protein Sci 18:180–195
16. Malisi C, Kohlbacher O, Höcker B (2009) Automated scaffold selection for enzyme design. Proteins 77:74–83
17. Lei Y, Luo W, Zhu Y (2011) A matching algorithm for catalytic residue site selection in computational enzyme design. Protein Sci 20:1566–1575
18. Zhang C, Lai L (2012) Automatch: target-binding protein design and enzyme design by automatic pinpointing potential active sites in available protein scaffolds. Proteins 80:1078–1094
19. Nosrati GR, Houk KN (2012) Saber: a computational method for identifying active sites for new reactions. Protein Sci 21:697–706
20. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D (2008) Kemp elimination catalysts by computational enzyme design. Nature 453: 190–195
21. Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, Hilvert D, Houk KN, Stoddard BL, Baker D (2008) De novo computational design of retro-Aldol enzymes. Science 319:1387–1391
22. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, Clair JLS, Gallaher JL, Hilvert D, Gelb MH, Stoddard BL, Houk KN, Michael FE, Baker D (2010) Computational design of an enzyme

catalyst for a stereoselective bimolecular Diels–Alder reaction. Science 329:309–313

23. Richter F, Blomberg R, Khare SD, Kiss G, Kuzin AP, Smith AJT, Gallaher J, Pianowski Z, Helgeson RC, Grjasnow A, Xiao R, Seetharaman J, Su M, Vorobiev S, Lew S, Forouhar F, Kornhaber GJ, Hunt JF, Montelione GT, Tong L, Houk KN, Hilvert D, Baker D (2012) Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. J Am Chem Soc 134:16197–16206
24. Bjelic S, Nivon LG, Celebi-Olcum N, Kiss G, Rosewall CF, Lovick HM, Ingalls EL, Gallaher JL, Seetharaman J, Lew S, Montelione GT, Hunt JF, Michael FE, Houk KN, Baker D (2013) Computational design of enone-binding proteins with catalytic activity for the Morita–Baylis–Hilman reaction. ACS Chem Biol 8: 749–757
25. Bolon DN, Mayo SL (2001) Enzyme-like proteins by computational design. Proc Natl Acad Sci USA 98:14274–14279
26. Privett HK, Kiss G, Lee TM, Blomberg R, Chica RA, Thomas LM, Hilvert D, Houk KN, Mayo SL (2012) Iterative approach to computational enzyme design. Proc Natl Acad Sci USA 109: 3790–3795
27. Canutescu AA, Dunbrack RL Jr (2003) Cyclic coordinate descent: a robotics algorithm for protein loop closure. Protein Sci 12:963–972
28. Coutsias EA, Seok C, Jacobson MP, Dill KA (2004) A kinematic view of loop closure. J Comput Chem 25:510–528
29. Tang K, Zhang J, Liang J (2014) Fast protein loop sampling and structure prediction using distance-guided sequential chain-growth Monte-Carlo method. PLoS Comput Biol 10:e1003539
30. Nocedal J, Wright SJ (1999) Numerical optimization. Springer, New York
31. Biegler LT, Grossmann IE, Westberg AW (1997) Systematic methods of chemical process design. Prentice Hall, Upper Saddle River
32. Facchinei F, Judice J, Soares J (1998) An active set Newton algorithm for large-scale nonlinear programs with box constraints. SIAM J Optim 8:158–186
33. Zhu Y (2007) Mixed-integer linear programming algorithm for a computational protein design problem. Ind Eng Chem Res 46:839–845
34. Huang X, Han K, Zhu Y (2013) Systematic optimization model and algorithm for binding sequence selection in computational enzyme design. Protein Sci 22:929–941
35. Tian Y, Huang X, Zhu Y (2015) Computational desing of enzyme-ligand binding using a combined energy function and deterministic sequence optimization algorithm. J Mol Model 21:191–204